Ausreißer

Autor und Herausgeber

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Zusammenfassung

Ein Ausreißer ist ein Datenpunkt, der so weit vom Rest eines Datensatz abweicht, dass vermutet wird, er sei auf andere Weise entstanden. Ein Ausreißer zu sein, ist ein Urteil, das gegen ein angegebenes Wahrscheinlichkeitsmodell oder Datensatz gefällt wird und keine Eigenschaft, die ein Datenpunkt von sich aus besitzt: Ein Tag ist typisch für das Jahr und extrem für seinen Monat, und zwei Merkmale können gemeinsam abweichen, während keines von beiden allein abweicht. Zwei Ursachen müssen unterschieden werden, eine Wahrscheinlichkeitsverteilung mit einem schweren Ausläufer und ein fehlerhafter oder kodierter Messwert, denn nur die zweite erfordert eine Korrektur der Daten. Bei mehr als zwei Merkmale macht die Dimensionsreduktion auf zwei gelernte Merkmale die abweichenden Datenpunkte in einem scatterplot sichtbar, weshalb dies ein üblicher erster Schritt ist. Wie stark ein Ausreißer die gelernte Hypothese verändert, hängt von der Verlustfunktion ab: Die squared error loss gewichtet ein Residuum mit seinem Quadrat und ist weitaus empfindlicher gegenüber einem falsch gemessenen Label als der absolute error loss. Die least absolute deviation regression, die Huber regression und die robuste Hauptkomponentenanalyse (robuste PCA) basieren auf diesem Unterschied, und ein contamination model gibt an, wie Ausreißer in die Stichprobe gelangen.

Definition

Eine Wetterstation in Krems an der Donau (Österreich) verzeichnete am 14. September 2024 81.7 Millimeter Regen, mehr als das Vierzigfache der durchschnittlichen Tagesmenge von 1.9 Millimetern in diesem Jahr. Dasselbe Archiv schreibt einen Tag ohne Niederschlag nicht als 0, sondern als -1 Millimeter, ein Code, den 199 der 366 Tage tragen. Die beiden Zahlen sind der größte und der kleinste Wert ihrer Spalte, und nur eine von ihnen kennzeichnet einen Tag, der einen zweiten Blick wert ist. Ein Ausreißer ist ein Datenpunkt, der so weit vom Rest eines Datensatz abweicht, dass vermutet wird, er sei auf andere Weise entstanden (Hawkins, 1980). Was als weit gilt, wird durch eine festgelegte Vorstellung von gewöhnlichem Verhalten bestimmt, daher ist ein Ausreißer ein Urteil, das gegen ein Wahrscheinlichkeitsmodell oder ein Datensatz gefällt wird und niemals eine Eigenschaft, die der Datenpunkt von sich aus besitzt (Chandola et al., 2009, Sect. 1.1).

Eine Änderung des Wahrscheinlichkeitsmodell ändert die Antwort, während die Daten unverändert bleiben. Man betrachte den Filter, der einen Tag als ungewöhnlich bezeichnet, wenn sein 15:00-Uhr-(UTC)-Messwert mehr als zwei Standardabweichungen vom Mittelwert abweicht. Unter einem Wahrscheinlichkeitsmodell, in dem dieser Messwert eine Gaußsche Zufallsvariable mit Mittelwert 16.3 und Standardabweichung 9.3 Grad ist, geschätzt aus den 365 Tagen des Jahres 2024, die ihn tragen, markiert der Filter zwei Tage. Mit dem Mittelwert und der Standardabweichung, die allein aus dem Kalendermonat des Tages geschätzt wurden, markiert er zwölf Tage, und nur den 9. Januar unter beiden. Am 3. Mai betrug der Messwert 12.7 Grad, was 0.39 Standardabweichungen unter dem Mittelwert des Jahres und 2.59 unter dem Mittelwert des Mai liegt: typisch für das Jahr, extrem für seinen Monat. Die −2.6 Grad des 8. Januar, der zweitkälteste Nachmittag des Jahres, verlaufen umgekehrt, 2.02 Standardabweichungen unter dem Mittelwert des Jahres und 1.47 unter dem Mittelwert des Januars, ein gewöhnlicher kalter Wintertag. Abb. 1 zeigt beide Filter über das Jahr.

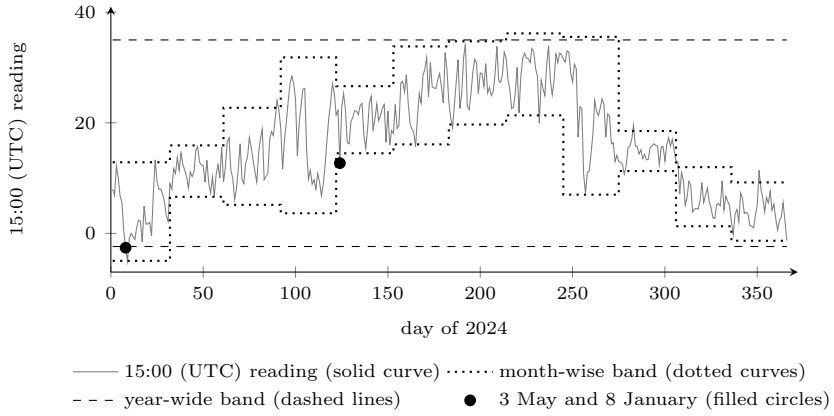

Abbildung 1: Die 15:00-Uhr-(UTC)-Lufttemperatur der 365 Tage des Jahres 2024 in Krems an der Donau, die sie tragen (durchgezogene Kurve), mit zwei Filtern, die einen Tag als ungewöhnlich bezeichnen, wenn der Messwert mehr als zwei Standardabweichungen vom Mittelwert abweicht: das Band des Wahrscheinlichkeitsmodell, geschätzt aus allen 365 Tagen (gestrichelte Linien), und das Band des Wahrscheinlichkeitsmodell, geschätzt aus dem Kalendermonat des Tages (gepunktete Kurven). Die gefüllten Kreise markieren den 3. Mai, innerhalb des jahresweiten Bandes und unterhalb des monatsweisen Bandes, und den 8. Januar, unterhalb des jahresweiten Bandes und innerhalb des monatsweisen Bandes. Daten generiert von `pythondemos/outlier.py`

Eine Abweichung kann auch in der Beziehung zwischen zwei Merkmale liegen und nicht in einem von beiden. Jeder Tag ist ein Datenpunkt mit dem Merkmalsvektor $\mathbf{x} = (x_1, x_2)^\top$, der seinen 06:00-Uhr-(UTC)- und seinen 15:00-Uhr-(UTC)-Messwert enthält. Im Jahr 2024 haben die beiden eine Korrelation

von 0.91, sodass ein Wahrscheinlichkeitsmodell, das $\mathbf{x}$ eine multivariate Normalverteilung mit Mittelwert $\boldsymbol{\mu}$ und Kovarianzmatrix $\mathbf{C}$ zuweist, erwartet, dass sie gemeinsam steigen und fallen. Es bewertet einen Tag anhand des quadrierten Mahalanobis-Abstands (Bishop, 2006, Sect. 2.3)

$$(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{C}^{-1} (\mathbf{x} - \boldsymbol{\mu}), \tag{1}$$

dessen Wahrscheinlichkeitsdichte der multivariate Normalverteilung eine abnehmende Funktion ist, sodass ein großer quadrierter Abstand eine kleine likelihood function bedeutet. Der größte Wert des quadrierten Mahalanobis-Abstands (1) im Jahr 2024, 9.0, gehört zum 4. März: ein 06:00-Uhr-Messwert von 0.7 und ein 15:00-Uhr-Messwert von 17.3 Grad, 1.16 und 0.11 Standardabweichungen von den jeweiligen Mittelwerte. Keines der Merkmal liegt allein jenseits der zwei Standardabweichungen des obigen Filters, und gemeinsam stellen sie einen Anstieg von 16.6 Grad von morgens bis nachmittags dar, den größten des Jahres.

Die Wetterstation zeichnet täglich acht Messwerte auf, und acht Zahlen sind mehr, als ein scatterplot zeigen kann. Die Dimensionsreduktion auf zwei Merkmale ist daher ein üblicher erster Schritt bei der Suche nach Ausreißern, da ein scatterplot zwei Zahlen pro Datenpunkt anzeigen kann. Die Hauptkomponentenanalyse (PCA) lernt eine lineare Abbildung von den acht standardisierten Messungen eines Tages auf $d' = 2$ Merkmale, die 63.6 Prozent ihrer Varianz behalten; die erste ist eine Temperaturrichtung und die zweite eine Feuchtigkeitsrichtung. In diesen beiden Merkmale sind die abweichenden Tage ohne jeden Filter sichtbar, wie Abb. 2 zeigt. Die Bewertung des quadrierten Mahalanobis-Abstands (1) in diesen beiden Merkmale anstatt in den beiden Temperaturen ergibt für den 14. September 50.7, gegenüber 22.9 für den zweitweitesten Tag und unter 8.3 für jeden der verbleibenden 363 Tage.

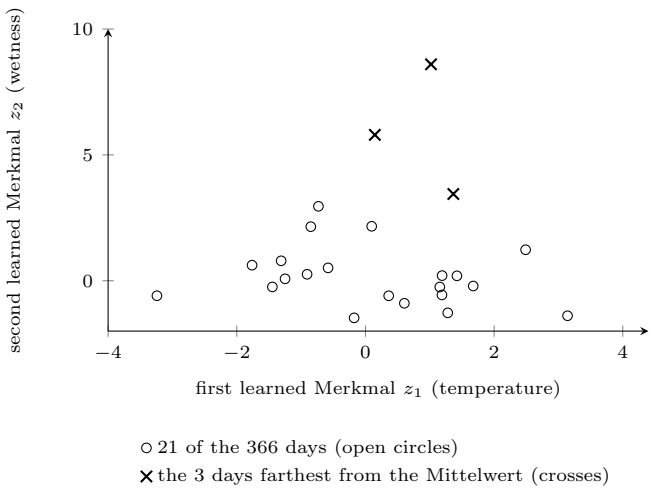

Abbildung 2: Jeder 18 der 366 Tage des Jahres 2024 in Krems, 21 Tage, platziert bei den zwei Merkmale, die PCA aus den acht standardisierten Messungen eines Tages lernt (offene Kreise), zusammen mit den 3 Tagen, die in diesen beiden Merkmale am weitesten vom Mittelwert entfernt sind (Kreuze). Der weiteste der drei ist der 14. September, der Tag mit 81.7 Millimetern Regen. Daten generiert von `pythondemos/outlier.py`

Der Code -1 der Einleitung ist von anderer Art. Das Archiv schreibt einen Tag ohne Niederschlag als -1 Millimeter und einen Tag mit weniger als 0.1 Millimetern als 0, sodass -1 überhaupt keine Menge ist. Ein Plausibilitätsfilter, der einen negativen Niederschlag ablehnt, markiert 199 der 366 Tage, 54 Prozent des Jahres, und keiner dieser Tage ist ungewöhnlich; als Menge gelesen, senkt der Code die jährliche Gesamtmenge von 712.0 auf 513.0 Millimeter. Die Abweichung rührt von der Annahme über die Kodierung der tabellarische Daten her, nicht vom Prozess, der die Daten erzeugt hat, und die Korrektur besteht darin, die Spalte neu zu kodieren, anstatt die Tage zu löschen. Die beiden Arten müssen unterschieden werden, bevor etwas entfernt wird, da ein Ausreißer entweder aus einer Wahrscheinlichkeitsverteilung mit einem schweren Ausläufer oder aus einem fehlerhaften Messwert (Bishop, 2006, Sect. 2.3.7) entsteht, und eine Kodierungskonvention ist keines von beidem.

Entfernung ist nicht die einzige Reaktion. Wie stark ein Ausreißer die gelernte Hypothese verändert, hängt von der Verlustfunktion ab: Die squared error loss gewichtet ein Residuum mit seinem Quadrat und der absolute error loss gewichtet es linear, was die least squares weitaus empfindlicher gegenüber einem grob falsch gemessenen Label macht als empirische Risikominimierung (ERM) mit dem absolute error loss (Hastie et al., 2009, Sect. 10.6). Bei den Krems-Daten lernt die least squares über das lineares Modell eine Hypothese, die den 15:00-Uhr-(UTC)-Messwert eines Tages aus seinem 06:00-Uhr-(UTC)-Messwert vorhersagt. Der 15:00-Uhr-Messwert vom 29. Dezember 2024 beträgt 5.5 Grad.

Ein falsch gesetztes Komma, das ihn als 55 aufzeichnet, verschiebt die Vorhersagen dieser Hypothese um bis zu 0.78 Grad und lässt die mit dem absolute error loss gelernte dort, wo sie war. Eine Erhöhung desselben Label auf 1000 verschiebt die least squares-Vorhersagen um 15.7 Grad, und die andere Hypothese bewegt sich immer noch überhaupt nicht. Wie weit ein einzelner ersetzter Datenpunkt die gelernte Hypothese verschieben kann, ist eine Frage der stability der Methode. Die least absolute deviation regression, die Huber regression mit der Huber loss und die robuste Hauptkomponentenanalyse (robuste PCA) für eine Merkmalsmatrix basieren auf diesem Unterschied, und ein contamination model gibt an, wie Ausreißer in die Stichprobe gelangen und was dann garantiert werden kann.

Siehe auch: anomaly detection, contamination model, robustness, stability, Huber regression, least absolute deviation regression, robuste PCA, absolute error loss, squared error loss, Wahrscheinlichkeitsmodell, Dimensionsreduktion, PCA.

Querverweise

Literatur

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.

2. Chandola et al. (2009). Anomaly Detection: A Survey. ACM Comput. Surv., 41(3), pp. 1–58.

3. Hawkins (1980). Identification of Outliers. Chapman and Hall, London.

4. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.